J. Appl. Prob. **47,** 732–751 (2010)
Printed in England

RECURRENCE EQUATIONS FOR THE PROBABILITY DISTRIBUTION OF SAMPLE CONFIGURATIONS IN EXACT POPULATION GENETICS MODELS

SABIN LESSARD,* *Université de Montréal*

Abstract

Recurrence equations for the number of types and the frequency of each type in a random sample drawn from a finite population undergoing discrete, nonoverlapping generations and reproducing according to the Cannings exchangeable model are deduced under the assumption of a mutation scheme with infinitely many types. The case of overlapping generations in discrete time is also considered. The equations are developed for the Wright–Fisher model and the Moran model, and extended to the case of the limit coalescent with nonrecurrent mutation as the population size goes to ∞ and the mutation rate to 0. Computations of the total variation distance for the distribution of the number of types in the sample suggest that the exact Moran model provides a better approximation for the sampling formula under the exact Wright–Fisher model than the Ewens sampling formula in the limit of the Kingman coalescent with nonrecurrent mutation. On the other hand, this model seems to provide a good approximation for a Λ-coalescent with nonrecurrent mutation as long as the probability of multiple mergers and the mutation rate are small enough.

Keywords: Ewens sampling formula; coalescent theory; Cannings model; Wright–Fisher model; Moran model; exchangeable population genetics model; Λ-coalescent

2010 Mathematics Subject Classification: Primary 60C05
Secondary 92D15

1. Introduction

The Ewens sampling formula (see Ewens (1972), with an addendum by Karlin and McGregor (1972), and see Ewens (1990), (2004, pp. 111–119) for historical perspectives) gives the joint distribution of the number of types with the number of copies of each type in a random sample taken from a large population under neutral mutation allowing for infinitely many possible types. This formula was first deduced under the assumption of a binomial reproduction scheme according to the Wright–Fisher model for a finite haploid population undergoing discrete, nonoverlapping generations as the population size N goes to ∞ and the mutation rate per individual per generation u goes to 0, such that the scaled mutation rate with N generations taken as the unit of time tends to a finite limit $\theta/2$. More generally, the formula applies to a sample drawn at random from a large population reproducing according to the Cannings exchangeable model in discrete time (see Cannings (1974)), as long as the limit coalescent backward in time as the population size N goes to ∞, with c_N^{-1} time steps taken as the unit of time, is the Kingman coalescent (see Kingman (1982)). Here c_N represents the probability that two individuals randomly chosen at the same time are copies of the same individual one

Received 16 February 2010; revision received 17 May 2010.

* Postal address: Département de Mathématiques et de Statistique, Université de Montréal, Montréal, Québec H3C 3J7, Canada. Email address: lessards@dms.umontreal.ca

time step back. This is known to be the case if and only if the rate of triple mergers tends to 0 (see Möhle (2000) and Möhle and Sagitov (2001)). This is the case, for instance, for the Moran model with overlapping generations with one individual replaced at a time (see Moran (1958)).

The Ewens sampling formula comes up in other biological contexts, e.g. in birth-and-death processes with immigration from a large mainland population (see Joyce and Tavaré (1987) and Tavaré (1989)) or in the Wright-island model for populations subdivided into a large number of demes (see Wright (1931) and Moran (1959)), with migration playing the role of mutation and identity-by-descent to a common ancestor defining types. When dealing with local populations, it is particularly relevant to relax the assumptions of large population size and small migration rate. This has been considered with local reproduction of the Wright–Fisher or Moran type (see Wakeley (2003), Wakeley and Takahashi (2004), and Fu (2006)) for instance, and exact sampling formulae for these cases have ascertained diffusion approximations for finitely-many-islands models as the number of demes goes to ∞ (see Lessard (2007)). The establishment of these formulae relied on a direct combinatorial approach previously used to deduce the Ewens sampling formula (see Griffiths and Lessard (2005), and see Hoppe (1984), (1987) for related arguments in Polya-like urn models). The approach has been extended to diploid populations (see Lessard (2009)).

In this paper we consider the exact Cannings model for a finite population and we deduce recurrence equations for the joint distribution of the number of types with the number of copies of each type in a random sample, as well as the marginal distribution for the number of types. We develop the equations in the cases of the Wright–Fisher model and the Moran model. Passing to the limit of large population size and small mutation rate yields recurrence equations in the case of a general coalescent as well as in the particular case of a Λ-coalescent, which allows for multiple mergers but not for simultaneous mergers, in agreement with previous studies (see Möhle (2006), and see Dong *et al.* (2007) and Freund and Möhle (2009) for more recent studies). The total variation distance for the distribution of the number of types is considered to compare the exact distribution in the case of the Wright–Fisher model to the exact distribution in the case of the Moran model and to the distribution predicted by the Ewens sampling formula. This distribution is also compared to the distribution obtained in the case of a Λ-coalescent modeling rare events of replacement of a fixed proportion of the population by the descendants of a single individual, as studied in Eldon and Wakeley (2006).

2. Recurrence equation for the probability of an ordered sample configuration

Consider a population of N haploid individuals undergoing discrete generations, and suppose that there are infinitely many possible types. Under neutrality, the N individuals of any given generation, arbitrarily labeled from 1 to N, are supposed to leave descendants in the next generation, possibly including some of the parents themselves, in numbers given by exchangeable random variables $z_1, \dots, z_N$, respectively, such that the total population size remains unchanged, that is, $z_1 + \cdots + z_N = N$ (see Cannings (1974)). From one generation to the next, mutation events creating entirely novel types are assumed to occur such that the number of mutant descendants left by the N parents, denoted by $\mu_1, \dots, \mu_N$, are also exchangeable random variables. Then, the number of nonmutant descendants left by the N parents, denoted by $\nu_1, \dots, \nu_N$, are exchangeable and given by $\nu_j = z_j - \mu_j$ for $j = 1, \dots, N$. More precisely, the N two-dimensional random variables $(\mu_1, \nu_1), \dots, (\mu_N, \nu_N)$, with $\mu_j + \nu_j = z_j$ for $j = 1, \dots, N$, are exchangeable.

Let $p(\boldsymbol{n})$, where $\boldsymbol{n} = (n_1, \dots, n_k)$ with $|\boldsymbol{n}| = \sum_{i=1}^{k} n_i = n$, be the stationary probability of a particular *ordered* sample configuration with multiplicities of types given by $\boldsymbol{n}$. More precisely, this is the probability that a sample of n individuals drawn at random without replacement in any given generation of the population at equilibrium exhibits k types, arbitrarily labeled from 1 to k with type i represented exactly n_i times at given positions for $i = 1, \dots, k$, once the sample is arbitrarily ordered. Note that

$$p^0(\boldsymbol{n}) = \frac{n!\, p(\boldsymbol{n})}{\prod_{i=1}^k n_i!} \tag{1}$$

is the corresponding probability for labeled types at any positions, and that

$$p^*(\boldsymbol{n}) = \frac{p^0(\boldsymbol{n})}{\prod_{j=1}^n b_j(\boldsymbol{n})!} \tag{2}$$

is the corresponding probability for unlabeled types at any positions. Here $b_j(\boldsymbol{n})$ represents the number of types represented j times in a sample with multiplicities of types given by $\boldsymbol{n} = (n_1, \dots, n_k)$, that is, the number of indices i such that $n_i = j$. For instance, $p^0(2, 1, 1)$ is the probability of all the ordered sample configurations $(1, 2, 2, 3)$, $(3, 2, 2, 1)$, $(2, 2, 1, 3)$, $(2, 2, 3, 1)$, $(1, 3, 2, 2)$, $(3, 1, 2, 2)$, $(2, 1, 3, 2)$, $(2, 3, 1, 2)$, $(1, 2, 3, 2)$, $(3, 2, 1, 2)$, $(2, 1, 2, 3)$, $(2, 3, 2, 1)$, each one being of probability $p(2, 1, 1)$, while $p^*(2, 1, 1) = p^0(2, 1, 1)/(2!\,1!\,1!)$ is the probability of three types in a sample of size 4, two types being represented once and one type being represented twice.

Condition on coalescence events involving lineages of sampled individuals of the same type and mutation events occurring on lineages associated with types represented once in the sample from one generation to the previous generation. Then the probability of a particular ordered sample configuration with multiplicities of types given by $\boldsymbol{n}$ can be expressed in the form

$$p(\boldsymbol{n}) = \sum_{\boldsymbol{0} \le \boldsymbol{l} \le \boldsymbol{n}} q(\boldsymbol{n}, \boldsymbol{l}) p(\boldsymbol{n} - \boldsymbol{l}), \tag{3}$$

with the convention that $p(\boldsymbol{0}) = 1$, where $\boldsymbol{l} = (l_1, \dots, l_k)$, with $0 \le l_i \le n_i$, but $l_i = n_i$ possible only if $n_i = 0$ or 1 for $i = 1, \dots, k$, gives the number of individuals of each type lost in one generation backward in time as a result of coalescence or mutation events. Therefore, $\boldsymbol{n} - \boldsymbol{l}$ gives the multiplicities of types in the ordered configuration of the parents of the nonmutant descendants. Note that $l = |\boldsymbol{l}| = \sum_{i=1}^k l_i$ is the total number of coalescence or mutation events, while $l_i = n_i = 1$ corresponds to the mutation event that gave rise to type i for $i = 1, \dots, k$. Moreover, $q(\boldsymbol{n}, \boldsymbol{l})$ represents the probability that n sampled individuals arranged in a specific order and partitioned into labeled subsets of sizes $n_1, \dots, n_k$, respectively, are exact copies of $n - l$ parents arranged in a specific order and partitioned into labeled subsets of sizes $n_1 - l_1, \dots, n_k - l_k$, respectively, with the convention that a mutant individual is not an exact copy of any parent (see Figure 1).

Explicitly, we have

$$q(\boldsymbol{n}, \boldsymbol{l}) = \binom{N}{\boldsymbol{n} - \boldsymbol{l}} \binom{N}{\boldsymbol{n}}^{-1} \sum_{(a_{i,r})} \mathrm{E}\left((M)_m \prod_{i=1}^k \prod_{r=1}^{n_i - l_i} \binom{\nu_{i,r}}{a_{i,r}} \right), \tag{4}$$

where $\nu_{i,r}$ stands for the number of nonmutant descendants left in the next generation by a parent of type i for $r = 1, \dots, n_i - l_i$, while $M = \sum_{j=1}^N \mu_j$ designates the total random number of mutant descendants among which m are sampled and labeled. In fact, $\nu_{i,r}$ can be replaced with

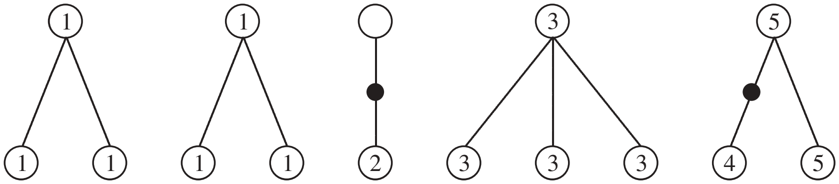

FIGURE 1: Ordered sample of $n = 10$ individuals exhibiting $k = 5$ types labeled from 1 to 5 with multiplicities $n_1 = 4, n_2 = 1, n_3 = 3, n_4 = 1, n_5 = 1$ in the individuals themselves and $n_1 - l_1 = 2$, $n_2 - l_2 = 0, n_3 - l_3 = 1, n_4 - l_4 = 0, n_5 - l_5 = 1$ in the $n - l = 4$ parents of the nonmutant individuals. A mutation event is indicated by a black circle on a line of descent.

$\nu_r + \sum_{j=1}^{i-1}(n_j - l_j)$, since the variables $\nu_1, \ldots, \nu_N$ are exchangeable, while m represents the number of types i such that $l_i = n_i = 1$. This may occur if and only if $n_i - l_i = 0$ under the constraint that $n_i = l_i$ is possible only if $n_i = 0$ or 1, so that

$$m = \sum_{\{i\,:\, n_i - l_i = 0\}} n_i.$$

Moreover, $(a_{i,r})$ designates an array of positive integers satisfying

$$a_{i,1} + \cdots + a_{i,n_i - l_i} = n_i$$

for all i such that $n_i - l_i \geq 1$. In (4) we used the notation

$$\binom{N}{\boldsymbol{n}} = \frac{N!}{(\prod_{i=1}^{k} n_i!)(N - n)!}$$

for the multinomial coefficient,

$$(M)_m = m!\binom{M}{m} = M(M-1)\cdots(M-m+1)$$

for the falling factorial, and the convention that

$$\prod_{r=1}^{n_i - l_i} \binom{\nu_{i,r}}{a_{i,r}} = 1$$

whenever $n_i - l_i = 0$.

Equation (4) can be deduced by ordering the N descendants of a given generation as follows: first order all the mutant descendants and then order the nonmutant descendants from the N parents in turn. The denominator in (4) is obtained by sampling in turn without replacement subsets of n_i descendants for $i = 1, \ldots, k$. The numerator in (4) is obtained by sampling in turn without replacement m mutant descendants and then subsets of the nonmutant descendants from $n_i - l_i$ parents, with at least one descendant per parent and a total of n_i nonmutant descendants for $i = 1, \ldots, k$. Moreover, the parents of subsets of nonmutant descendants are all different.

Note that, using an inclusion–exclusion argument and the exchangeability property as in Möhle (2004), the probability of no mutation or coalescence event affecting the sampled individuals from one generation to the previous generation, which depends on the sample size n, is given by

$$q(\boldsymbol{n},\boldsymbol{0}) = \mathrm{E}\left(\prod_{r=1}^{n} \nu_r\right) = \sum_{j=1}^{n} (-1)^{n-j}\binom{n}{j}\mathrm{E}\left(\binom{\sum_{r=1}^{j}\nu_r}{n}\right). \tag{5}$$

Substituting (5) into (3), rearranging terms, and defining

$$Q(\boldsymbol{n},\boldsymbol{l}) = \frac{q(\boldsymbol{n},\boldsymbol{l})}{1-q(\boldsymbol{n},\boldsymbol{0})}, \tag{6}$$

which is a conditional probability given that at least one event of coalescence or mutation affects sampled individuals from one generation to the previous generation, we obtain the following result.

Result 1. *The probability of a particular ordered sample configuration in the Cannings model with infinitely many types satisfies the recurrence equation*

$$p(\boldsymbol{n}) = \sum_{\boldsymbol{0}<\boldsymbol{l}\le\boldsymbol{n}} Q(\boldsymbol{n},\boldsymbol{l})p(\boldsymbol{n}-\boldsymbol{l}). \tag{7}$$

The summation is over $\boldsymbol{l} = (l_1,\dots,l_k)$, $0 \le l_i \le n_i$ *for* $i = 1,\dots,k$, *with a strict inequality on the left-hand side for some* i *such that* $l = \sum_{i=1}^{k} l_i > 0$ *and an equality possible on the right-hand side only if* $n_i = 0$ *or* 1.

Corollary 1. *Considering all generations backward in time with at least one event of coalescence or mutation affecting the ancestry of the original sample, we have*

$$p(\boldsymbol{n}) = \sum_{\tau^*=1}^{n}\sum_{(\boldsymbol{l}_\tau)}\prod_{\tau=1}^{\tau^*} Q\left(\boldsymbol{n}-\sum_{\sigma=1}^{\tau-1}\boldsymbol{l}_\sigma,\boldsymbol{l}_\tau\right).$$

The summation is over a number of steps backward in time τ^* *going from 1 to n and over all histories of coalescence or mutation events represented by a sequence* $(\boldsymbol{l}_\tau)$. *The sequence* $(\boldsymbol{l}_\tau)$ *is compatible with an ordered sample configuration with multiplicities of types given by* $\boldsymbol{n}$, *that is, satisfying* $0 < \boldsymbol{l}_\tau \le \boldsymbol{n} - \sum_{\sigma=1}^{\tau-1}\boldsymbol{l}_\sigma$ *as in Result 1 for* $\tau = 1,\dots,\tau^*$, *with* $\sum_{\sigma=1}^{\tau^*}\boldsymbol{l}_\sigma = \boldsymbol{n}$.

Remark. Substituting (1), (2), (4), and (6) into recurrence equation (7), and using the identity $M = N - \sum_{j=1}^{N}\nu_j$ and the exchangeability property of $\nu_1,\dots,\nu_N$, it has been checked that

$$\sum_{\substack{n_1\ge\cdots\ge n_k\ge 0\\ n_1+\cdots+n_k=n}} p^*(\boldsymbol{n}) = 1 \tag{8}$$

for $n = 2, 3, 4$ from explicit expressions for $p^*(\boldsymbol{n})$ (see Appendix A).

3. Recurrence equation for the number of types in the sample

The probability that a sample of size n contains exactly k different types for $1 \le k \le n$ can be expressed as

$$f_n(k) = \sum_{\substack{n_1\ge\cdots\ge n_k\ge 1\\ n_1+\cdots+n_k=n}} p^*(\boldsymbol{n}). \tag{9}$$

Therefore, a recurrence equation for $f_n(k)$ could be deduced from the recurrence equation for $p(\boldsymbol{n})$ given in Result 1, since $p^*(\boldsymbol{n})$ is related to $p(\boldsymbol{n})$ through (1) and (2).

An alternative approach is to resort to arguments similar to those previously used. Noting that the number of mutation events is necessarily less than or equal to k with equality possible only if $k = n$, while the number of coalescence events is always bounded by $n - k$, we obtain the recurrence equation

$$f_n(k) = \sum_{m=0}^{k-1+\delta_{n,k}} \sum_{c=0}^{n-k} h(n, m, c) f_{n-m-c}(k - m), \tag{10}$$

with the convention that $f_0(0) = 1$, and $\delta_{n,k} = 1$ if $k = n$ and 0 otherwise. Here $h(n, m, c)$ is the probability of m mutation events and c coalescence events without mutation in a sample of size n from one generation to the previous generation. This probability is given explicitly by

$$h(n, m, c) = \binom{N}{n-l}\binom{N}{n}^{-1} \sum_{(a_r)} \mathrm{E}\left(\binom{M}{m} \prod_{r=1}^{n-l} \binom{\nu_r}{a_r}\right), \tag{11}$$

where $l = m + c$ and (a_r) is an array of $n - l$ positive integers satisfying

$$a_1 + \cdots + a_{n-l} = n - m.$$

When $m = n$ and $c = 0$, we have

$$h(n, n, 0) = \mathrm{E}\left(\binom{M}{m}\right)\binom{N}{n}^{-1}.$$

Let us define

$$H(n, m, c) = \frac{h(n, m, c)}{1 - h(n, 0, 0)}, \tag{12}$$

where $h(n, 0, 0) = \mathrm{E}(\prod_{r=1}^{n} \nu_r)$, which is a conditional probability given that at least one event of coalescence or mutation affects the sample from one generation to the previous generation. Then, rearranging the terms in (10) yields the following result.

Result 2. *The probability for the number of types in a sample under the Cannings model with infinitely many types satisfies the recurrence equation*

$$f_n(k) = \sum_{m=0}^{k-1+\delta_{n,k}} \sum_{c=\delta_{m,0}}^{n-k} H(n, m, c) f_{n-m-c}(k - m), \tag{13}$$

where δ designates the Kronecker delta.

Corollary 2. *Considering all generations backward in time with at least one event of coalescence or mutation affecting the ancestry of the original sample, we obtain*

$$f_n(k) = \sum_{\tau^*=1}^{n} \sum_{(m_\tau, c_\tau)} \prod_{\tau=1}^{\tau^*} H\left(n - \sum_{\sigma=1}^{\tau-1} m_\sigma - \sum_{\sigma=1}^{\tau-1} c_\sigma, m_\tau, c_\tau\right),$$

where $m_\tau \geq 0$ and $c_\tau \geq 0$, $\tau = 1, \ldots, \tau^*$, with at least one strict inequality, with $m_{\tau^*} > 0$, $c_{\tau^*} = 0$, $\sum_{\tau=1}^{\tau^*} m_\tau = k$, and $\sum_{\tau=1}^{\tau^*} c_\tau = n - k$.

Remark. Explicit expressions for $f_n(k)$ can be obtained recursively by substituting (11) and (12) into (13), or, equivalently, by using (9). Then, owing to (8), the identity $\sum_{k=1}^{n} f_n(k) = 1$ is ascertained at least for $n = 2, 3, 4$.

4. Nonoverlapping generations

In this section we consider the case of nonoverlapping generations with independent mutation events creating entirely novel types for descendants with probability u per descendant. The numbers of nonmutant descendants left by the N parents of a given generation, denoted by $\nu_1, \dots, \nu_N$, given the total numbers of descendants left by the N parents, denoted by $z_1, \dots, z_N$, which are assumed to be exchangeable, are independent random variables that follow a binomial distribution with parameters $z_1, \dots, z_N$, respectively, and $1 - u$.

Considering n parents in a given generation and conditioning on the total numbers of descendants that they leave in the next generation, we obtain

$$\mathrm{E}\biggl(\prod_{r=1}^{n} \nu_r\biggr) = (1-u)^n \, \mathrm{E}\biggl(\prod_{r=1}^{n} z_r\biggr).$$

More generally, we have

$$\mathrm{E}\biggl((M)_m \prod_{r=1}^{n-l} (\nu_r)_{a_r}\biggr) = u^m (1-u)^{n-m} (N-n+m)_m \, \mathrm{E}\biggl(\prod_{r=1}^{n-l} (z_r)_{a_r}\biggr),$$

with the notation $(x)_r = x(x-1)\cdots(x-r+1)$ for the falling factorial, where $a_1, \dots, a_{n-l}$ are positive integers satisfying $\sum_{r=1}^{n-l} a_r = n - m$. This can be checked by considering the conditional probability generating function

$$G(\boldsymbol{s}, \boldsymbol{t} \mid \boldsymbol{z}) = \mathrm{E}\biggl(\prod_{j=1}^{N} s_j^{\mu_j} t_j^{\nu_j} \biggm| z_1, \dots, z_N\biggr) = \prod_{j=1}^{N} (s_j u + t_j(1-u))^{z_j},$$

where $\boldsymbol{s} = (s_1, \dots, s_N)$, $\boldsymbol{t} = (t_1, \dots, t_N)$, and $\boldsymbol{z} = (z_1, \dots, z_N)$. Then, the partial derivative $\partial^n G(s\mathbf{1}, \boldsymbol{t} \mid \boldsymbol{z}) / \partial s^m \partial t_1^{a_1} \cdots \partial t_{n-l}^{a_{n-l}}$, where $\mathbf{1} = (1, \dots, 1)$ is the N-dimensional unit vector, evaluated at $s = 1$ and $\boldsymbol{t} = \mathbf{1}$, yields

$$\mathrm{E}\biggl((M)_m \prod_{r=1}^{n-l} (\nu_r)_{a_r} \biggm| z_1, \dots, z_N\biggr) = u^m (1-u)^{n-m} (N-n+m)_m \prod_{r=1}^{n-l} (z_r)_{a_r}. \tag{14}$$

Taking the expected value with respect to $\boldsymbol{z}$ in (14) gives the desired result. Similarly, we have

$$\mathrm{E}\biggl((M)_m \prod_{i=1}^{k} \prod_{r=1}^{n_i - l_i} (\nu_{i,r})_{a_{i,r}}\biggr) = u^m (1-u)^{n-m} (N-n+m)_m \times \mathrm{E}\biggl(\prod_{i=1}^{k} \prod_{r=1}^{n_i - l_i} (z_{i,r})_{a_{i,r}}\biggr),$$

under the condition that $\sum_{i=1}^{k} \sum_{r=1}^{n_i - l_i} a_{i,r} = n - m$. This is the case for the array $(a_{i,r})$ in (4).

Finally, using the identities $(N-n+m)_m(N-n)! = (N-n+m)!$ and $\prod_{i=1}^k n_i! = \prod_{i=1}^k (n_i - m_i)!$, where $m_i = 1$ if $n_i = l_i = 1$ and 0 otherwise, we obtain

$$h(n,m,c) = \frac{\binom{N}{n-l}\binom{n}{m} u^m (1-u)^{n-m} \sum_{(a_r)} \mathrm{E}\big(\prod_{r=1}^{n-l} \binom{z_r}{a_r}\big)}{\binom{N}{n-m}}$$

for the probability in (11) and

$$q(\boldsymbol{n},\boldsymbol{l}) = \frac{\binom{N}{\boldsymbol{n}-\boldsymbol{l}} u^m (1-u)^{n-m} \sum_{(a_{i,r})} \mathrm{E}\big(\prod_{i=1}^k \prod_{r=1}^{n_i-l_i} \binom{z_{i,r}}{a_{i,r}}\big)}{\binom{N}{\boldsymbol{n}-\boldsymbol{m}}}$$

for the probability in (4), where $\boldsymbol{m} = (m_1, \dots, m_k)$ and $m = \sum_{i=1}^k m_i$. Therefore, we obtain the following result.

Result 3. *In the case of nonoverlapping generations with infinitely many types, the coefficients in the recurrence equations of Results 1 and 2 are given by*

$$Q(\boldsymbol{n},\boldsymbol{l}) = \frac{\binom{N}{\boldsymbol{n}-\boldsymbol{l}} u^m (1-u)^{n-m} \sum_{(a_{i,r})} \mathrm{E}\big(\prod_{i=1}^k \prod_{r=1}^{n_i-l_i} \binom{z_{i,r}}{a_{i,r}}\big)}{\binom{N}{\boldsymbol{n}-\boldsymbol{m}}\big(1-(1-u)^n \,\mathrm{E}\big(\prod_{r=1}^n z_r\big)\big)}$$

and

$$H(n,m,c) = \frac{\binom{N}{n-l}\binom{n}{m} u^m (1-u)^{n-m} \sum_{(a_r)} \mathrm{E}\big(\prod_{r=1}^{n-l} \binom{z_r}{a_r}\big)}{\binom{N}{n-m}\big(1-(1-u)^n \,\mathrm{E}\big(\prod_{r=1}^n z_r\big)\big)},$$

respectively, where $l = m + c$.

Remark. The coefficients in Result 3 can be obtained directly by conditioning on all the mutation events affecting sampled descendants and the total numbers of descendants left by their parents.

5. Wright–Fisher model

In the case of nonoverlapping generations with a reproduction scheme according to the Wright–Fisher model (see Fisher (1930, pp. 83–96) and Wright (1931)), the random vector $(z_1, \dots, z_N)$ for each generation follows a multinomial distribution with parameters N and $(1/N, \dots, 1/N)$, whose probability generating function is

$$G(\boldsymbol{s}) = \mathrm{E}\bigg(\prod_{j=1}^N s_j^{z_j}\bigg) = \bigg(\frac{1}{N}\sum_{j=1}^N s_j\bigg)^N,$$

where $\boldsymbol{s} = (s_1, \dots, s_N)$. Therefore, we obtain

$$\mathrm{E}\bigg(\prod_{r=1}^n z_r\bigg) = \frac{\partial^n}{\partial s_1 \cdots \partial s_n} G(\boldsymbol{s})\bigg|_{\boldsymbol{s}=\boldsymbol{1}} = \frac{(N)_n}{N^n}.$$

More generally, we have

$$\mathrm{E}\bigg(\prod_{r=1}^{n-l} (z_r)_{a_r}\bigg) = \frac{\partial^n}{\partial s_1^{a_1} \cdots \partial s_{n-l}^{a_{n-l}}} G(\boldsymbol{s})\bigg|_{\boldsymbol{s}=\boldsymbol{1}} = \frac{(N)_{n-m}}{N^{n-m}}, \tag{15}$$

if $a_1, \dots, a_{n-l}$ are positive integers satisfying $\sum_{r=1}^{n-l} a_r = n - m$, and similarly

$$\mathrm{E}\left(\prod_{i=1}^{k}\prod_{r=1}^{n_i-l_i}(z_{i,r})_{a_{i,r}}\right) = \frac{(N)_{n-m}}{N^{n-m}}, \tag{16}$$

if $a_{i,r} \geq 1$ for all (i, r) with $1 \leq r \leq n_i - l_i$ and $\sum_{i=1}^{k}\sum_{r=1}^{n_i-l_i} a_{i,r} = n - m$. On the other hand, we have

$$\frac{1}{(n-l)!}\sum_{(a_r)}\frac{(n-m)!}{\prod_{r=1}^{n-l} a_r!} = S_{n-m}^{(n-l)}$$

and

$$\sum_{(a_{i,r})}\frac{(\prod_{i=1}^{k}(n_i)_{l_i})}{(\prod_{i=1}^{k}\prod_{r=1}^{n_i-l_i} a_{i,r}!)} = \prod_{i=1}^{k} S_{n_i}^{(n_i-l_i)},$$

where $S_n^{(r)}$ denotes the Stirling number of the second kind, which represents the number of ways that a set of n distinct elements can be partitioned into r nonempty subsets (see, e.g. Abramowitz and Stegun (1965, p. 824)). Substituting the above expressions and the identity $(N)_{n-m}(N - n + m)! = (N)_{n-l}(N - n + l)!$ into Result 3 yields the following result.

Result 4. *In the case of the Wright–Fisher model with infinitely many types, the coefficients in the recurrence equations of Results 1 and 2 are given by*

$$Q(\boldsymbol{n}, \boldsymbol{l}) = \frac{(Nu)^m(1-u)^{n-m}(N)_{n-l}\prod_{i=1}^{k} S_{n_i}^{(n_i-l_i)}}{N^n - (1-u)^n(N)_n}$$

and

$$H(n, m, c) = \binom{n}{m}\frac{(Nu)^m(1-u)^{n-m}(N)_{n-l}S_{n-m}^{(n-l)}}{N^n - (1-u)^n(N)_n}, \tag{17}$$

respectively, where $l = m + c$.

6. Moran model

In the case of overlapping generations in discrete time, the exchangeable random variables $z_1, \dots, z_N$ with $z_1 + \cdots + z_N = N$ represent the numbers of descendants left by the N individuals of the population in one time step, including the individuals that survive. Assuming that a parent leaving any new descendant survives, that only the new descendants can mutate, and that all mutation events are independent and each one occurs with the same probability u, the only difference with the case of nonoverlapping generations is that the variable $\nu_r - 1$ given z_r for $z_r \geq 1$ follows a binomial distribution with parameters $z_r - 1$ and $1 - u$ for $r = 1, \dots, N$.

In the case of the Moran model (see Moran (1958), (1962, pp. 78–85)), the vector

$$z = (z_1, z_2, z_3, \dots, z_N)$$

is a random permutation of $(2, 0, 1, \dots, 1)$ with probability $(N-1)/N$, and $(1, \dots, 1)$ otherwise. This models a population of size N in which, at each time step, an individual is chosen at random to produce one offspring and this offspring replaces one of the individuals of the population chosen at random, not excluding the parent of the offspring. Moreover, it is assumed that the offspring produced can mutate with probability u independently of everything else.

Conditioning on the distribution of z, we find that

$$\mathrm{E}\biggl(\prod_{r=1}^{n}\nu_r\biggr) = 1 - \frac{n(n-1+u)}{N^2} - \frac{n(N-n)u}{N^2},$$

since the variable $\prod_{r=1}^{n}\nu_r$ takes the value 0 with probability $(n(N-1+u))/N^2$, the value 2 with probability $(n(N-n)(1-u))/N^2$, and the value 1 otherwise. Moreover, only one event of coalescence or mutation can occur at a time, so that

$$Q(\boldsymbol{n},\boldsymbol{l}) = 0$$

unless $\boldsymbol{l} = \boldsymbol{e}_j$, the standard jth unit vector (that is, $l_j = 1$ and $l_i = 0$ for all $i \neq j$) for some $1 \leq j \leq k$. In this case, if $n_j = 1$ then there is mutation of type j one generation backward in time and no other mutation event ($m = 1$) or coalescence event involving other types ($a_{i,r} = 1$ for all $i \neq j$), so that

$$\mathrm{E}\biggl((M)_m\prod_{i=1}^{k}\prod_{r=1}^{n_i-l_i}(\nu_{i,r})_{a_{i,r}}\biggr) = \mathrm{E}\biggl(M\prod_{r=1}^{n-1}\nu_r\biggr) = \frac{(N-n+1)u}{N},$$

since $M\prod_{r=1}^{n-1}\nu_r$ takes the value 1 with probability $((N-n+1)u)/N$ and the value 0 otherwise. On the other hand, if $\boldsymbol{l} = \boldsymbol{e}_j$ with $n_j > 1$ then there is coalescence of two lineages of type j one generation backward in time and no other coalescence event ($a_{i,r} = 2$ for one and only one r for $k = j$ and 1 otherwise) or mutation event ($m = 0$), so that

$$\begin{aligned}\mathrm{E}\biggl((M)_m\prod_{i=1}^{k}\prod_{r=1}^{n_i-l_i}(\nu_{i,r})_{a_{i,r}}\biggr) &= \mathrm{E}\biggl(\nu_1(\nu_1-1)\prod_{r=2}^{n-1}\nu_r\biggr)\\ &= \frac{2(N-n+1)(1-u)}{N^2},\end{aligned}$$

since $\nu_1(\nu_1-1)\prod_{r=2}^{n-1}\nu_r$ takes the value 2 with probability $((N-n+1)(1-u))/N^2$ and the value 0 otherwise.

Moreover, we note that there are $n_j - 1$ arrays $(a_{i,r})$ satisfying the above conditions. In such circumstances, (4) and (6) lead to

$$Q(\boldsymbol{n},\boldsymbol{e}_j) = \frac{n_j N u}{n(n-1+(N-n+1)u)}$$

if $n_j = 1$ and

$$Q(\boldsymbol{n},\boldsymbol{e}_j) = \frac{n_j(n_j-1)(1-u)}{n(n-1+(N-n+1)u)}$$

if $n_j > 1$.

Multiplying all terms for a history of coalescence or mutation events represented by a sequence of standard unit vectors $(\boldsymbol{l}_\tau)$ compatible with an ordered sample configuration with multiplicities given by $\boldsymbol{n}$, that is, satisfying $0 < \boldsymbol{l}_\tau \leq \boldsymbol{n} - \sum_{\sigma=1}^{\tau-1}\boldsymbol{l}_\sigma$ for $\tau = 1,\dots,\tau^*$, with $\sum_{\sigma=1}^{\tau^*}\boldsymbol{l}_\sigma = \boldsymbol{n}$, we obtain

$$\frac{(\prod_{i=1}^{k}n_i!)(\prod_{i=1}^{k}(n_i-1)!)N^k u^k(1-u)^{n-k}}{n!\,(\prod_{r=1}^{n}(r-1+(N-r+1)u))}$$

for $\prod_{\tau=1}^{\tau^*} Q(\boldsymbol{n} - \sum_{\sigma=1}^{\tau-1} \boldsymbol{l}_\sigma, \boldsymbol{l}_\tau)$ if $\tau^* = n$, and 0 otherwise. Moreover, the number of possible histories with $\tau^* = n$ is $n!/(\prod_{i=1}^{k} n_i!)$. Therefore, introducing the scaled mutation rate θ defined as

$$\theta = \frac{Nu}{1-u}$$

and the notation

$$\theta^{(n)} = \prod_{r=1}^{n} (\theta + r - 1)$$

for the rising factorial, Corollary 1 leads to the following result.

Result 5. *The probability of a particular ordered sample configuration in the Moran model in discrete time with infinitely many types is given by*

$$p(\boldsymbol{n}) = \frac{\theta^k \prod_{i=1}^{k} (n_i - 1)!}{\theta^{(n)}}. \tag{18}$$

Remark. Result 5 is in agreement with Trajstman (1974), who deduced the stationary distribution in the whole population assuming that the offspring produced at each time step can replace any other individual chosen at random, including the parent, with θ defined as $Nu/(1-u)$. See, e.g. Ewens (2004, pp. 118, 340) for a discussion and some perspectives.

Remark. For the number of types in the sample, (1), (2), (9), and (18) yield

$$f_n(k) = \frac{\theta^k |s_n^{(k)}|}{\theta^{(n)}}, \tag{19}$$

where

$$|s_n^{(k)}| = \sum_{\substack{n_1 \geq \cdots \geq n_k \geq 1 \\ n_1 + \cdots + n_k = n}} \frac{n!}{(\prod_{i=1}^{k} n_i)(\prod_{j=1}^{n} b_j(\boldsymbol{n})!)}$$

is the unsigned Stirling number of the first kind that represents the number of permutations of n elements which contain exactly k permutation cycles. This is a well-known formula (see, e.g. Ewens (2004, pp. 114, 118)). Note that, since $\sum_{k=1}^{n} f_n(k) = 1$, it follows from (19) that $|s_n^{(k)}|$ is the coefficient of θ^k in $\theta^{(n)}$, which is also a well-known fact (see, e.g. Abramowitz and Stegun (1965, p. 824)).

7. Limit coalescent

We come back to the Cannings model with nonoverlapping generations. Let us define

$$\Phi((a_r)) = \frac{(N)_{n-l} \,\mathrm{E}(\prod_{r=1}^{n-l} (z_r)_{a_r})}{(N)_{n-m}}, \tag{20}$$

where $a_1, \ldots, a_{n-l} \geq 1$ and $a_1 + \cdots + a_{n-l} = n - m$. This represents the probability of a *particular* (a_r) merger one generation backward in time. More precisely, this is the probability that $n - m$ individuals chosen at random without replacement in a given generation descend from exactly $n - l$ parents in the previous generation, labeled arbitrarily from 1 to $n - l$, with exactly a_r of the individuals in particular descending from parent r for $r = 1, \ldots, n - l$.

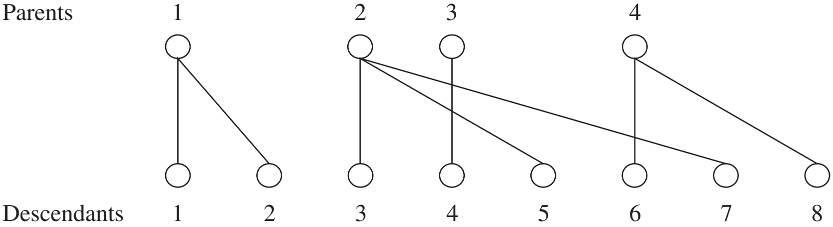

FIGURE 2: Eight nonmutant individuals labeled with the ranks, from 1 to 8, in which they are sampled without replacement in a given generation, and partitioned into subsets $S_1 = \{1, 2\}$, $S_2 = \{3, 5, 7\}$, $S_3 = \{4\}$, and $S_4 = \{6, 8\}$ according to their parents labeled with the ranks, from 1 to 4, in which they appear for the first time in sampling the eight descendants.

Expression (20) can be explained as follows. Let us sample $n - m$ individuals one at a time at random without replacement in a given generation. This procedure will order not only the sampled individuals but also their *parents*, which are both labeled with the ranks in which they appear for the first time (see Figure 2). Regrouping the sampled individuals descending from the same parents gives a partition of $\{1, \dots, n - m\}$ into subsets $S_1, \dots, S_{n-l}$ for some $l \geq m$ with S_r containing the descendants of parent r among the $n - m$ sampled individuals, for $r = 1, \dots, n - l$. Given that the total numbers of descendants left by these parents are $z_1, \dots, z_{n-l}$, the probability of a partition satisfying $|S_r| = a_r \geq 1$ for $r = 1, \dots, n - l$ with $\sum_{r=1}^{n-l} a_r = n - m$ is $\prod_{j=1}^{n-m} p_j$, where

$$p_j = \frac{(N - r + 1)z_r}{N - j + 1}$$

if j is the smallest integer in S_r (the first descendant of parent r) and

$$p_j = \frac{z_r - \alpha + 1}{N - j + 1}$$

if j is the αth smallest integer in S_r (the αth descendant of parent r). Multiplying all terms and taking the expectation over $z_1, \dots, z_{n-l}$ gives (20).

Similarly, the expression

$$\Phi((a_{i,r})) = \frac{(N)_{n-l}\,\mathrm{E}(\prod_{i=1}^{k}\prod_{r=1}^{n_i-l_i}(z_{i,r})_{a_{i,r}})}{(N)_{n-m}},$$

where $a_{i,r} \geq 1$ for $r = 1, \dots, n_i - l_i$ and $i = 1, \dots, k$, with

$$\sum_{i=1}^{k}(n_i - l_i) = n - l, \qquad \sum_{r=1}^{n_i-l_i} a_{i,r} = n_i, \qquad \text{and} \qquad \sum_{i=1}^{k}\sum_{r=1}^{n_i-l_i} a_{i,r} = n - m,$$

represents the probability of a particular $(a_{i,r})$ merger one generation backward in time. This is consistent with the previous definition since $(a_{i,r})$ is an array of $n - l$ positive integers that sum up to $n - l$. Conversely, the previous definition corresponds to the case of an array $(a_{i,r})$ with $k = 1$, $n_1 = n - m$, and $n_1 - l_1 = n - l$.

Moreover, note that

$$\Phi(\mathbf{1}) = \mathrm{E}\bigg(\prod_{r=1}^{n} z_r\bigg),$$

where $\mathbf{1}$ stands for an array (a_r) defined by $a_r = 1$ for $r = 1, \ldots, n$, gives the probability of no merger affecting a sample of size n one generation backward in time.

Let

$$c_N = \Phi(2) = \frac{\mathrm{E}(z_1(z_1 - 1))}{N - 1},$$

where the 2 stands for an array (a_r) defined by $a_r = 2$ for $r = 1$. This is the probability that two individuals chosen at random in a given generation descend from the same parent. Suppose that this probability goes to 0 in the limit of a large population size, that is, $\lim_{N\to\infty} c_N = 0$. Assume that

$$\lim_{N\to\infty} \frac{\Phi((a_{i,r}))}{c_N} = \phi((a_{i,r})) < \infty,$$

as soon as $(a_{i,r})$ is different from $\mathbf{1}$ (see, e.g. Möhle and Sagitov (2001)). This limit corresponds to the rate of a particular $(a_{i,r})$ merger with c_N^{-1} generations taken as the unit of time when N goes to ∞. Moreover, let

$$\lim_{N\to\infty} \frac{1 - \Phi(\mathbf{1})}{c_N} = \lambda_n < \infty.$$

This is the limiting rate of change of n lineages by coalescence. In particular, we have $\lambda_2 = \phi(2) = 1$. Note that $\lim_{N\to\infty} \Phi(\mathbf{1}) = 1$, under the above assumptions. Finally, assume that $\lim_{N\to\infty} 2uc_N^{-1} = \theta$. Then, letting N go to ∞ in Result 3 leads to the following result.

Result 6. *As $N \to \infty$, the coefficients in the recurrence equation of Result 1 in the case of nonoverlapping generations with infinitely many types as given in Result 3 tend to*

$$Q_\infty(\boldsymbol{n}, \boldsymbol{l}) = \frac{\prod_{i=1}^{k} (n_i)_{l_i}}{n\theta/2 + \lambda_n} \sum_{(a_{i,r})} \frac{\phi((a_{i,r}))}{\prod_{i=1}^{k} \prod_{r=1}^{n_i - l_i} a_{i,r}!} \tag{21}$$

if $\boldsymbol{m} = \mathbf{0}$ and $\boldsymbol{l} > \mathbf{0}$,

$$Q_\infty(\boldsymbol{n}, \boldsymbol{l}) = \frac{\theta/2}{n\theta/2 + \lambda_n}$$

if $\boldsymbol{l} = \boldsymbol{m} = \boldsymbol{e}_i$ for some $i = 1, \ldots, k$, and 0 otherwise. Similarly, the coefficients in Result 2 tend to

$$H_\infty(n, m, c) = \frac{(n)_c}{n\theta/2 + \lambda_n} \sum_{(a_r)} \frac{\phi((a_r))}{\prod_{r=1}^{n-c} a_r!} \tag{22}$$

if $m = 0$ and $c > 0$,

$$H_\infty(n, m, c) = \frac{n\theta/2}{n\theta/2 + \lambda_n}$$

if $m = 1$ and $c = 0$, and 0 otherwise.

Remark. Result 6 implies that mutation events are isolated in the case of the limit coalescent: mutation events are not compatible with coalescence events and only one mutation event can occur at a time.

Remark. Explicit expressions for $p^*(\boldsymbol{n})$, $n = 2, 3, 4$, in the limit coalescent obtained from Result 6 are given in Appendix B. It can be checked that these probabilities sum up to 1 for a fixed sample size. Result 6 is in agreement with Freund and Möhle (2009), but it was obtained here as a limit case of an exact result for a finite population.

8. The Λ-coalescent

In the case of no simultaneous mergers in the limit coalescent, called the Λ-coalescent (see Pitman (1999) and Sagitov (1999)), only the arrays of numbers that have only one element larger than 1, given by $c + 1$, have to be considered in (21) and (22) of Result 6. This implies that there are $n_i - c$ possible arrays $(a_{i,r})$ in (21) with $\boldsymbol{l} = c\boldsymbol{e}_i$ for some $i = 1, \dots, k$, and $n - c$ possible arrays (a_r) in (22), with both arrays satisfying

$$\prod_{i=1}^{k} \prod_{r=1}^{n_i - l_i} a_{i,r}! = \prod_{r=1}^{n-c} a_r! = (c+1)!.$$

Moreover, there exists a nonnegative finite measure Λ on [0, 1] such that

$$\phi((a_{i,r})) = \phi((a_r)) = \lambda_{n,c+1} = \int_{[0,1]} x^{c-1}(1-x)^{n-c-1}\Lambda(\mathrm{d}x)$$

and

$$\lambda_n = \sum_{c=1}^{n-1} \binom{n}{c+1} \lambda_{n,c+1} = \int_{[0,1]} \frac{1-(1-x)^{n-1}(1-x+nx)}{x^2} \Lambda(\mathrm{d}x)$$

for $n \geq 2$. All this leads to the following result, which is in agreement with Möhle (2006) (see also Dong *et al.* (2007) for more general conditions on the coalescent process for regenerative recursions).

Result 7. *In the case of a Λ-coalescent in the limit as $N \to \infty$, the limit coefficients in Result 6 are given by*

$$Q_\infty(\boldsymbol{n}, c\boldsymbol{e}_i) = \binom{n_i}{c+1} \frac{\lambda_{n,c+1}}{n\theta/2 + \lambda_n}$$

if $1 \leq c \leq n_i - 1$,

$$Q_\infty(\boldsymbol{n}, \boldsymbol{e}_i) = \frac{\theta/2}{n\theta/2 + \lambda_n}$$

if $n_i = 1$ *for some* $i = 1, \dots, k$, *and* $Q_\infty(\boldsymbol{n}, \boldsymbol{l}) = 0$ *otherwise, and*

$$H_\infty(n, 0, c) = \binom{n}{c+1} \frac{\lambda_{n,c+1}}{n\theta/2 + \lambda_n} \tag{23}$$

if $1 \leq c \leq n - 1$,

$$H_\infty(n, 1, 0) = \frac{n\theta/2}{n\theta/2 + \lambda_n}, \tag{24}$$

and $H_\infty(n, m, c) = 0$ *otherwise.*

Remark. Kingman's (1982) coalescent corresponds to the case $\Lambda = \delta_0$, where δ_0 denotes the Dirac measure at 0. Then, we have $\lambda_{n,c+1} = 1$ when $c = 1 \geq n - 1$, and 0 otherwise, while $\lambda_n = n(n-1)/2$. This is the case in the limit of a large population in the Wright–Fisher model owing to (15) and (16), with $c_N = N^{-1}$, and, therefore, $\theta = \lim_{N\to\infty} 2Nu$. In the expression given in Corollary 1 for the probability of a particular ordered sample with multiplicities of types given by $\boldsymbol{n}$, only the histories $(\boldsymbol{l}_\tau)$ with $\boldsymbol{l}_\tau = \boldsymbol{e}_i$ for some $i = 1, \dots, k$ and all $\tau = 1, \dots, \tau^*$, and, therefore, for $\tau^* = n$ since $\sum_{\tau=1}^{\tau^*} \boldsymbol{l}_\tau = \boldsymbol{n}$, have to be considered, for which

$$\prod_{\tau=1}^{\tau^*} Q\biggl(\boldsymbol{n} - \sum_{\sigma=1}^{\tau-1} \boldsymbol{l}_\sigma, \boldsymbol{l}_\tau\biggr) = \frac{\theta^k (\prod_{i=1}^k n_i!)(\prod_{i=1}^k (n_i - 1)!)}{n!\theta^{(n)}},$$

owing to Result 7. Moreover, there are $n!\,/(\prod_{i=1}^k n_i!)$ possible histories, from which $p(\boldsymbol{n})$ takes the form of (18) as in the Moran model, in agreement with Ewens's (1972) sampling formula. As a consequence, the formula for the number of types will also be the same and given by (19). Note that this formula can be deduced directly from (23) and (24) using the fact that $|s_n^{(k)}|$ is the coefficient of θ^k in $\theta^{(n)}$.

Remark. If $\Lambda = \delta_\psi$ for some $0 < \psi < 1$ (the Dirac coalescent) then we have

$$\lambda_{n,c+1} = \psi^{c-1}(1-\psi)^{n-c-1} \tag{25}$$

for $1 \le c \le n-1$ and

$$\lambda_n = \frac{1 - (1-\psi)^n - n\psi(1-\psi)^{n-1}}{\psi^2} \tag{26}$$

for $n \ge 2$. Such a Λ-coalescent models a situation where mergers occur when the offspring of a single individual count for a proportion ψ of the population. More precisely, this situation is obtained in the limit of a large population size when, e.g. in each generation with probability $N^{-\alpha}$ for some $0 < \alpha < 1$, a single individual chosen at random in the population has a probability ψ of being the parent of each descendant in the next generation compared to $(1 - \psi)/(N-1)$ for each of the other individuals. Otherwise, this probability is $1/N$ for every individual. This is a modified Wright–Fisher model, which reduces to the standard Wright–Fisher model when $\psi = 1/N$. In this case,

$$c_N = \psi^2 N^{-\alpha}(1 + O(N^{1-\alpha})), \tag{27}$$

and the scaled mutation rate is defined as $\theta = \lim_{N\to\infty} 2uc_N^{-1}$. It was applied to oyster populations in Eldon and Wakeley (2006).

9. Total variation distance

Being valid in the realm of the Kingman coalescent with mutation, the Ewens sampling formula provides a good approximation for samples taken from a finite population obeying the Cannings model as long as the population size is large enough, the mutation rate small enough, and the distribution of the number of descendants of an individual not too skewed. In order to evaluate to what extent the approximation is accurate when these conditions are relaxed, the total variation distance for the distribution of the number of types can be considered.

Let us denote by $f_n^{\mathrm{WF}}(k)$, $f_n^{\mathrm{M}}(k)$, $f_n^{\mathrm{K}}(k)$, and $f_n^{\mathrm{EW}}(k)$ the probability of k types in a random sample of size n taken from a population under the assumptions of the exact Wright–Fisher (WF) model ((13) and (17)), the exact Moran (M) model ((19) with $\theta = 2Nu/(1-u)$), the

Table 1: Total variation distance between different models for the probability distribution of the number of types in a random sample of size n from a population of size N with mutation rate u per generation, or θ per unit of time.

n	N	u	K − WF	M − WF	θ	ψ	K − EW
5	10	0.1	0.0626	0.0259	2	0.1	0.0250
		0.3	0.1845	0.0749		0.3	0.0778
		0.5	0.3200	0.1089		0.5	0.1343
		0.7	0.3869	0.0976		0.7	0.2109
		0.9	0.3922	0.0437		0.9	0.3233
	50	0.1	0.0534	0.0195	10	0.1	0.0477
		0.3	0.1030	0.0326		0.3	0.1453
		0.5	0.1131	0.0304		0.5	0.2412
		0.7	0.1111	0.0223		0.7	0.3284
		0.9	0.1015	0.0090		0.9	0.3994
	100	0.1	0.0429	0.0148	20	0.1	0.0439
		0.3	0.0601	0.0183		0.3	0.1230
		0.5	0.0613	0.0159		0.5	0.1862
		0.7	0.0584	0.0113		0.7	0.2342
		0.9	0.0526	0.0045		0.9	0.2775
10	10	0.1	0.1054	0.0489	2	0.1	0.0572
		0.3	0.3152	0.1310		0.3	0.1601
		0.5	0.4952	0.2014		0.5	0.2618
		0.7	0.6411	0.2682		0.7	0.3414
		0.9	0.8409	0.1754		0.9	0.4221
	50	0.1	0.0908	0.0338	10	0.1	0.0902
		0.3	0.1930	0.0736		0.3	0.2653
		0.5	0.3214	0.1034		0.5	0.4107
		0.7	0.3764	0.0900		0.7	0.5583
		0.9	0.3743	0.0396		0.9	0.6798
	100	0.1	0.0768	0.0274	20	0.1	0.1001
		0.3	0.1699	0.0574		0.3	0.3257
		0.5	0.2170	0.0620		0.5	0.5129
		0.7	0.2267	0.0483		0.7	0.6383
		0.9	0.2133	0.0201		0.9	0.7112

Kingman (K) coalescent ((19) with $\theta = 2Nu$), and the Eldon and Wakeley (EW) Λ-coalescent ((13) with (23), (24), (25), (26), (27), and $\theta = \lim_{N\to\infty} 2uc_N^{-1}$), respectively.

The total variation distance between two models, say K and WF, for a sample size n is

$$dv_n(\mathrm{K}, \mathrm{WF}) = \frac{1}{2}\sum_{k=1}^{n} |f_n^{\mathrm{K}}(k) - f_n^{\mathrm{WF}}(k)|.$$

Of course, the total variation distance depends not only on the sample size but also on N and u, or θ. Numerical results for a range of parameter values are presented in Table 1.

The Moran model with adjusted population size and mutation rate generally provides a better approximation to the exact Wright–Fisher model than the Kingman coalescent with the population size taken as the unit of time for the sample sizes ($n = 5$ or 10), population sizes ($N = 10, 50$, or 100), and mutation rates per generation ($u = 0.1, 0.3, 0.5, 0.7$, or 0.9) that have been considered. Moreover, in both cases, the total variation distance increases with the sample size, which is expected since the sample size corresponds to the number of possible values for the number of types. It also generally decreases with increasing population size, as Table 1 suggests, but there are exceptions. For instance, the value of $dv_5(\mathrm{K}, \mathrm{M})$ in the case $u = 0.1$ goes from 0.019 96 when $N = 30$ to 0.020 13 when $N = 40$. Finally, in both cases, the total variation distance increases as the mutation rate increases at least for small values of the mutation rate, which is expected since then simultaneous mutation events are rare, but it often decreases for large enough values of the mutation rate, which is surprising.

The total variation distance between the Eldon and Wakeley Λ-coalescent and the Kingman coalescent increases with the sample size ($n = 5$ or 10) and the proportion of the population replaced ($\psi = 0.1, 0.3, 0.5, 0.7$, or 0.9), but not necessarily with the scaled mutation rate ($\theta = 2$, 10, or 20). The values of this rate correspond to $N = 10, 50$, or 100, respectively, with $u = 0.1$ in the case of the Wright–Fisher model. Table 1 shows that $dv_n(\mathrm{K}, \mathrm{EW})$ is of the same order of magnitude as $dv_n(\mathrm{K}, \mathrm{WF})$ in the $\psi = 0.1$ case, but much greater when ψ is larger.

Appendix A

From (1), (2), (4), and (6), Result 1 yields

$$\begin{aligned}
\tilde{p}(1) &= N(1 - \mathrm{E}(\nu_1)),\\
\tilde{p}(2) &= N\,\mathrm{E}(\nu_1(\nu_1 - 1)),\\
\tilde{p}(1, 1) &= 2N\,\mathrm{E}(M\nu_1) + \mathrm{E}(M(M - 1)),\\
\tilde{p}(3) &= 3N(N - 1)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2)p^*(2) + N\,\mathrm{E}(\nu_1(\nu_1 - 1)(\nu_1 - 2)),\\
\tilde{p}(2, 1) &= 3N(N - 1)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2)p^*(1, 1) + 3N(N - 1)\,\mathrm{E}(M\nu_1\nu_2)p^*(2)\\
&\quad + 3N\,\mathrm{E}(M\nu_1(\nu_1 - 1)),\\
\tilde{p}(1, 1, 1) &= 3N(N - 1)\,\mathrm{E}(M\nu_1\nu_2)p^*(1, 1) + 3N\,\mathrm{E}(M(M - 1)\nu_1)\\
&\quad + \mathrm{E}(M(M - 1)(M - 2)),\\
\tilde{p}(4) &= 6N(N - 1)(N - 2)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2\nu_3)p^*(3)\\
&\quad + 3N(N - 1)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2(\nu_2 - 1))p^*(2)\\
&\quad + 4N(N - 1)\,\mathrm{E}(\nu_1(\nu_1 - 1)(\nu_1 - 2)\nu_2)p^*(2)\\
&\quad + N\,\mathrm{E}(\nu_1(\nu_1 - 1)(\nu_1 - 2)(\nu_1 - 3)),\\
\tilde{p}(3, 1) &= 4N(N - 1)(N - 2)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2\nu_3)p^*(2, 1)\\
&\quad + 4N(N - 1)(N - 2)\,\mathrm{E}(M\nu_1\nu_2\nu_3)p^*(3)\\
&\quad + 4N(N - 1)\,\mathrm{E}(\nu_1(\nu_1 - 1)(\nu_1 - 2)\nu_2)p^*(1, 1)\\
&\quad + 12N(N - 1)\,\mathrm{E}(M\nu_1(\nu_1 - 1)\nu_2)p^*(2) + 4N\,\mathrm{E}(M\nu_1(\nu_1 - 1)(\nu_1 - 2)),\\
\tilde{p}(2, 2) &= 2N(N - 1)(N - 2)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2\nu_3)p^*(2, 1)\\
&\quad + 3N(N - 1)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2(\nu_2 - 1))p^*(1, 1),\\
\tilde{p}(2, 1, 1) &= 6N(N - 1)(N - 2)\,\mathrm{E}(\nu_1(\nu_1 - 1)\nu_2\nu_3)p^*(1, 1, 1)\\
&\quad + 4N(N - 1)(N - 2)\,\mathrm{E}(M\nu_1\nu_2\nu_3)p^*(2, 1)\\
&\quad + 12N(N - 1)\,\mathrm{E}(M\nu_1(\nu_1 - 1)\nu_2)p^*(1, 1)\\
&\quad + 6N(N - 1)\,\mathrm{E}(M(M - 1)\nu_1\nu_2)p^*(2) + 6N\,\mathrm{E}(M(M - 1)\nu_1(\nu_1 - 1)),
\end{aligned}$$

$$\begin{aligned}\tilde{p}(1,1,1,1) = {} & 4N(N-1)(N-2)\,\mathrm{E}(M\nu_1\nu_2\nu_3)p^*(1,1,1)\\ & + 6N(N-1)\,\mathrm{E}(M(M-1)\nu_1\nu_2)p^*(1,1) + 4N\,\mathrm{E}(M(M-1)(M-2)\nu_1)\\ & + \mathrm{E}(M(M-1)(M-2)(M-3)),\end{aligned}$$

where

$$\tilde{p}(\boldsymbol{n}) = (N)_n\biggl(1 - \mathrm{E}\biggl(\prod_{r=1}^{n}\nu_r\biggr)\biggr)p^*(\boldsymbol{n}),$$

with $\nu_1, \dots, \nu_N$ being exchangeable nonnegative integer-valued random variables and $M = N - \sum_{j=1}^{N}\nu_j \geq 0$. Then, we find that

$$\begin{aligned}\tilde{p}(2) + \tilde{p}(1,1) &= N\,\mathrm{E}(\nu_1(\nu_1-1)) + 2N\,\mathrm{E}(M\nu_1) + \mathrm{E}(M(M-1))\\ &= N(N-1)(1-\mathrm{E}(\nu_1\nu_2)),\\ \tilde{p}(3) + \tilde{p}(2,1) + \tilde{p}(1,1,1) &= N\,\mathrm{E}(\nu_1(\nu_1-1)(\nu_1-2)) + 3N(N-1)\,\mathrm{E}(\nu_1(\nu_1-1)\nu_2)\\ &\quad + 3N\,\mathrm{E}(M\nu_1(\nu_1-1)) + 3N(N-1)\,\mathrm{E}(M\nu_1\nu_2)\\ &\quad + 3N\,\mathrm{E}(M(M-1)\nu_1) + \mathrm{E}(M(M-1)(M-2))\\ &= N(N-1)(N-2)(1-\mathrm{E}(\nu_1\nu_2\nu_3)),\\ \tilde{p}(4) + \tilde{p}(3,1) + \tilde{p}(2,2) + {}& \tilde{p}(2,1,1) + \tilde{p}(1,1,1,1)\\ &= 6N(N-1)(N-2)\,\mathrm{E}(\nu_1(\nu_1-1)\nu_2\nu_3)\\ &\quad + 3N(N-1)\,\mathrm{E}(\nu_1(\nu_1-1)\nu_2(\nu_2-1))\\ &\quad + 4N(N-1)\,\mathrm{E}(\nu_1(\nu_1-1)(\nu_1-2)\nu_2)\\ &\quad + N\,\mathrm{E}(\nu_1(\nu_1-1)(\nu_1-2)(\nu_1-3))\\ &\quad + 4N(N-1)(N-2)\,\mathrm{E}(M\nu_1\nu_2\nu_3)\\ &\quad + 12N(N-1)\,\mathrm{E}(M\nu_1(\nu_1-1)\nu_2)\\ &\quad + 4N\,\mathrm{E}(M\nu_1(\nu_1-1)(\nu_1-2))\\ &\quad + 6N(N-1)\,\mathrm{E}(M(M-1)\nu_1\nu_2)\\ &\quad + 6N\,\mathrm{E}(M(M-1)\nu_1(\nu_1-1))\\ &\quad + 4N\,\mathrm{E}(M(M-1)(M-2)\nu_1)\\ &\quad + \mathrm{E}(M(M-1)(M-2)(M-3))\\ &= N(N-1)(N-2)(N-3)(1-\mathrm{E}(\nu_1\nu_2\nu_3\nu_4)).\end{aligned}$$

Appendix B

Using (1) and (2), Result 6 yields

$$\begin{aligned}p^*(2) &= \frac{\phi(2)}{\theta+\lambda_2},\\ p^*(1,1) &= \frac{\theta}{\theta+\lambda_2},\\ p^*(3) &= \frac{2\phi(3) + 6\phi(2,1)p^*(2)}{3\theta+2\lambda_3},\\ p^*(2,1) &= \frac{3\theta p^*(2) + 6\phi(2,1)p^*(1,1)}{3\theta+2\lambda_3},\end{aligned}$$

$$
\begin{aligned}
p^*(1,1,1) &= \frac{3\theta p^*(1,1)}{3\theta + 2\lambda_3},\\
p^*(4) &= \frac{\phi(4) + 6\phi(2,1,1)p^*(3) + 3\phi(2,2)p^*(2) + 4\phi(3,1)p^*(2)}{2\theta + \lambda_4},\\
p^*(3,1) &= \frac{4\phi(2,1,1)p^*(2,1) + 4\phi(3,1)p^*(1,1) + 2\theta p^*(3)}{2\theta + \lambda_4},\\
p^*(2,2) &= \frac{2\phi(2,1,1)p^*(2,1) + 3\phi(2,2)p^*(1,1)}{2\theta + \lambda_4},\\
p^*(2,1,1) &= \frac{6\phi(2,1,1)p^*(1,1,1) + 2\theta p^*(2,1)}{2\theta + \lambda_4},\\
p^*(1,1,1,1) &= \frac{2\theta p^*(1,1,1)}{2\theta + \lambda_4},
\end{aligned}
$$

with $\lambda_2 = \phi(2)$, $\lambda_3 = \phi(3) + 3\phi(2,1)$, and $\lambda_4 = \phi(4) + 4\phi(3,1) + 3\phi(2,2) + 6\phi(2,1,1)$.

Acknowledgements

The author is grateful to an anonymous referee for a careful reading of the manuscript and constructive comments. Research supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

Abramowitz, M. and Stegun, I. A. (eds) (1965). *Handbook of Mathematical Functions*. Dover, New York.

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv. Appl. Prob.* **6,** 260–290.

Dong, R., Gnedin, A. and Pitman, J. (2007). Exchangeable partitions derived from Markovian coalescents. *Ann. Appl. Prob.* **17,** 1172–1201.

Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* **172,** 2621–2633.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3,** 87–112.

Ewens, W. J. (1990). Population genetics theory—the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, ed. S. Lessard, Kluwer, Dordrecht, pp. 177–227.

Ewens, W. J. (2004). *Mathematical Population Genetics*. I. 2nd edn. Springer, New York.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Freund, F. and Möhle, M. (2009). On the number of allelic types for samples taken from exchangeable coalescents with mutation. *Adv. Appl. Prob.* **41,** 1082–1101.

Fu, Y. X. (2006). Exact coalescent for the Wright–Fisher model. *Theoret. Pop. Biol.* **69,** 385–394.

Griffiths, R. C. and Lessard, S. (2005). Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theoret. Pop. Biol.* **68,** 167–177.

Hoppe, F. M. (1984). Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.* **20,** 91–94.

Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25,** 123–159.

Joyce, P. and Tavaré, S. (1987). Cycles, permutations and the structures of the Yule process with immigration. *Stoch. Process. Appl.* **25,** 309–314.

Karlin, S. and McGregor, J. (1972). Addendum to a paper of W. Ewens. *Theoret. Pop. Biol.* **3,** 113–116.

Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.

Lessard, S. (2007). An exact sampling formula for the Wright–Fisher model and a solution to a conjecture about the finite-island model. *Genetics* **177,** 1249–1254.

Lessard, S. (2009). Diffusion approximations for one-locus multi-allele kin selection, mutation and random drift in group-structured populations: a unifying approach to selection models in population genetics. *J. Math. Biol.* **59,** 659–696.

Möhle, M. (2000). Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Adv. Appl. Prob.* **32,** 983–993.

Möhle, M. (2004). The time back to the most recent common ancestor in exchangeable population models. *Adv. Appl. Prob.* **36,** 78–97.

Möhle, M. (2006). On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli* **12,** 35–53.

Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Prob.* **29,** 1547–1562.

Moran, P. A. P. (1958). Random processes in genetics. *Proc. Camb. Phil. Soc.* **54,** 60–71.

Moran, P. A. P. (1959). The theory of some genetical effects of population subdivision. *Austral. J. Biol. Sci.* **12,** 109–116.

Moran, P. A. P. (1962). *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Prob.* **27,** 1870–1902.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Prob.* **36,** 1116–1125.

Tavaré, S. (1989). The genealogy of the birth, death, and immigration process. In *Mathematical Evolutionary Theory*, ed. M. W. Feldman, Princeton University Press, pp. 41–56.

Trajstman, A. C. (1974). On a conjecture of G. A. Watterson. *Adv. Appl. Prob.* **6,** 489–493.

Wakeley, J. (2003). Polymorphism and divergence for island-model species. *Genetics* **163,** 411–420.

Wakeley, J. and Takahashi, T. (2004). The many-demes limit for selection and drift in a subdivided population. *Theoret. Pop. Biol.* **66,** 83–91.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16,** 97–159.